An Identity of Chernoff Bounds with an Interpretation in Statistical Physics and Applications in Information Theory

Neri Merhav

March 16, 2008

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel

Abstract

An identity between two versions of the rate function of the probability a certain large deviations event, is established. This identity has an interpretation in statistical physics, namely, an isothermal equilibrium of a composite system that consists of multiple subsystems of particles. Several information–theoretic application examples, where the analysis of this large deviations probability naturally arises, are then described from the viewpoint of this statistical mechanical interpretation. This results in several relationships between information theory and statistical physics, which we hope, the reader will find insightful.

Index Terms: Large deviations theory, Chernoff bound, statistical physics, thermal equilibrium, equipartition, thermodynamics, phase transitions.

1 Introduction

Relationships between information theory and statistical physics have been extensively recognized over the last few decades, and they are drawn from many different aspects. We mention here only a few of them.

One such aspect is characterized by identifying structures of optimization problems pertaining to certain information–theoretic settings as being analogous to parallel structures that arise in statistical physics, and then borrowing statistical–mechanical insights, as well as powerful analysis techniques (like the replica method) from statistical physics to the dual information–theoretic

setting of interest. A very partial list of works along this line includes [1], [15], [16], [21], [22], [23] [24], [25], [34] (and references therein), [35], [36], [42], [43], [47], [48], [49], [50], and [51].

Another aspect pertains to the philosophy and the application of the maximum entropy principle, which emerged in statistical mechanics in the nineteenth century and has been advocated during the previous century in a wide variety of more general contexts, by Jaynes [18],[19],[20], and by Shore and Johnson [46], as a general guiding principle to problems in information theory (see, e.g., [7, Chap. 12] and references therein) and other areas, such as signal processing, in particular, speech coding (see, e.g., [13]) spectrum estimation (see, e.g., [5]), and others.

Yet another aspect is related to ideas and theories that underly the notion of 'trading' between information bits and energy, or heat. In particular, Landauer's erasure principle [27] is argued to provide a powerful link between information theory and physics and to suggest a physical theory of information (comprehensive overviews are included in, e.g., [30] and [41]). According to Landauer's principle, the erasure of every bit of information increases the thermodynamic entropy of the world by $k \ln 2$, where k is Boltzmann's constant, and so, information is actually physical.

Finally, shifting the focus into the direction of this paper's subject, we should mention the aspect of the interface between statistical physics and large deviations theory, a line of research advocated most prominently by Ellis [10],[11], and developed also by Oono [39], McAllester [31], and others. The main theme here evolves around the identification of Chernoff bounds and more general large deviations rate functions with free energies (along with their related partition functions), thermodynamical entropies, and the underlying maximum–entropy/equilibrium principle associated with them. In particular, Ellis' book [10] is devoted largely to the application of large deviations theory to the statistical physics pertaining to models of magnetic materials, like Ising spin glasses and others, in order to explore phase transitions phenomena of spontaneous magnetization (see also [34]).

This paper, which is mostly expository in character, lies in the intersection of information theory, large deviations theory, and statistical physics. In particular, we establish a simple identity between two quantities as they can both be interpreted as the rate function of a certain large deviations event that involves multiple distributions of sets of independent random variables (as opposed to the usual, single set of i.i.d. random variables). The analysis of this large deviations event is of

a general form that is frequently encountered in numerous applications in information theory (cf. Section 4). Its informal description is as follows: Let $v_1, \ldots, v_n$ be an arbitrary (deterministic) sequence whose components take on values in a finite set $\mathcal{V}$, and let $U_1, \ldots, U_n$ be a sequence of random variables where each component is generated independently according to a distribution $q(u_i|v_i)$, $i = 1, \ldots, n$. For a given function f and a constant E, we are interested in the large deviations analysis (Chernoff bound) of the probability of the event

$$\sum_{i=1}^{n} f(U_i, v_i) \leq nE, \tag{1}$$

assuming that the relative frequencies of the various symbols in $(v_1, \ldots, v_n)$ stabilize as n grows without bound, and assuming that E is sufficiently small to make this a rare event for large n.

There are (at least) two ways to derive a Chernoff bound on the probability of this event. The first is to treat the entire sequence of RV's, $\{f(U_i, v_i)\}$ as a whole, and the second is to partition it according to the various symbols $\{v_i\}$, i.e., to consider the separate large deviations events of the partial sums, $\sum_{i:v_i=v} f(U_i, v)$, $v \in \mathcal{V}$, for all possible allocations of the total 'budget' nE among the various $\{v\}$. These two approaches lead to two (seemingly) different expressions of Chernoff bounds, but since they are both exponentially tight, they must agree.

As will be described and discussed in Section 2, the identity between these two Chernoff bounds has a natural interpretation in statistical physics: it is viewed as a situation of thermal equilibrium (maximum entropy) in a system that consists of several subsystems (which can be of different kinds), each of them with many particles.

As will be shown in Section 4, the above–described problem of large deviations analysis of the event (1) is encountered in many applications in information theory, such as rate–distortion coding, channel capacity, hypothesis testing (signal detection, in particular), and many others which are not reviewed in this paper (cf. e.g. denoising and filtering under an excess loss probability criterion [40]). The above mentioned statistical mechanical interpretation then applies to all of them. Accordingly, Section 4 is devoted to expository descriptions of each of these applications, along with the underlying physics that is inspired by the proposed thermal equilibrium interpretation. The reader is assumed to have very elementary background in statistical physics.

The remaining part of this paper is organized as follows. In Section 2, we establish some notation conventions. In Section 3, we provide some very basic bacground in statistical physics,

which will be need for our derivations. In Section 4, we assert and prove our main result, which is the identity between the above described Chernoff bounds. Finally, in Section 5, we explore the application examples.

2 Notation

Throughout this paper, scalar random variables (RV's) will be denoted by the capital letters, like U,V,X, and Y, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with same symbols superscripted by the dimension. Thus, for example, X^n will denote a random n-vector $(X_1, \ldots, X_n)$, and $x^n = (x_1, ..., x_n)$ is a specific vector value in $\mathcal{X}^n$, the n-th Cartesian power of $\mathcal{X}$. The notations x_i^j and X_i^j, where i and j are integers and $i \leq j$, will designate segments $(x_i, \ldots, x_j)$ and $(X_i, \ldots, X_j)$, respectively, where for $i = 1$, the subscript will be omitted (as above). Sequences without specifying indices are denoted by $\{\cdot\}$. Sources and channels will be denoted generically by the letter P or Q. Specific letter probabilities corresponding to a source P will be denoted by the corresponding lower case letter, e.g., $p(v)$ is the probability of a letter $v \in \mathcal{V}$. A similar convention will be applied to a channel Q and the corresponding transition probabilities, e.g., $q(u|v)$, $u \in \mathcal{U}$, $v \in \mathcal{V}$. The cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$. Information theoretic quantities like entropies, and mutual informations will be denoted following the usual conventions of the information theory literature.

Notation pertaining to statistical physics will also follow, wherever possible, the customary conventions. I.e., k will denote Boltzmann's constant ($k = 1.38065 \times 10^{-23}$ Joules per Kelvin degree), T – absolute temperature (in Kelvin degrees), $\beta = 1/(kT)$ – the inverse temperature (in units of Joule^{-1}), E – energy, the letter Z will be used to denote partition functions, etc.

3 Background

In this section, we provide a brief account of the very basic background in statistical physics, which is needed for this paper.

Consider a physical system with a large number n of particles, which can be in a variety of

microscopic states ('microstates'), defined by combinations of, e.g., positions, momenta, angular momenta, spins, etc., of all n particles. For each such microstate of the system, which we shall designate by a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, there is an associated energy, given by an Hamiltonian (energy function), $\mathcal{E}(\boldsymbol{x})$. For example, if $x_i = (\boldsymbol{p}_i, h_i)$, where $\boldsymbol{p}_i$ is the momentum vector of particle number i and h_i is its height, then classically, $\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{n}[\frac{\|\boldsymbol{p}_i\|^2}{2m} + mgh_i]$, where m is the mass of each particle and g is the gravitation constant.

One of the most fundamental results in statistical physics (based on the law of energy conservation and the basic postulate that all microstates of the same energy level are equiprobable) is that when the system is in thermal equilibrium with its environment, the probability of a microstate $\boldsymbol{x}$ is given by the *Boltzmann–Gibbs* distribution

$$P(\boldsymbol{x}) = \frac{e^{-\beta\mathcal{E}(\boldsymbol{x})}}{Z(\beta)} \tag{2}$$

where as defined above, $\beta = 1/(kT)$, T being temperature, and $Z(\beta)$ is the normalization constant, called the *partition function*, which is given by

$$Z(\beta) = \sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}(\boldsymbol{x})}$$

or

$$Z(\beta) = \int d\boldsymbol{x} e^{-\beta\mathcal{E}(\boldsymbol{x})},$$

depending on whether $\boldsymbol{x}$ is discrete or continuous. The role of the partition function is by far deeper than just being a normalization factor, as it is actually the key quantity from which many macroscopic physical quantities can be derived, for example, the free energy[1] is $-\frac{1}{\beta}\ln Z(\beta)$, the average internal energy (i.e., the expectation of $\mathcal{E}(\boldsymbol{x})$ where $\boldsymbol{x}$ drawn is according (2)) is given by the negative derivative of $\ln Z(\beta)$, the heat capacity is obtained from the second derivative, etc. One of the ways to obtain eq. (2), is as the maximum entropy distribution under an energy constraint (owing to the second law of thermodynamics), where β plays the role of a Lagrange multiplier that controls this energy level.

Under certain assumptions on the Hamiltonian function, the following relations are well–known to hold and can be found in any textbook on elementary statistical physics (e.g.,[26]): Defining the

[1]The free energy means the maximum work that the system can carry out in any process of fixed temperature. The maximum is obtained when the process is reversible (slow, quasi–static changes in the system).

per–particle entropy, $S(E)$, associated with per–particle energy $E = \mathcal{E}(\boldsymbol{x})/n$, as $\lim_{n\to\infty}[\ln N(E)]/n$, [2] where $N(E)$ is the number of microstates $\{\boldsymbol{x}\}$ with energy level $\mathcal{E}(\boldsymbol{x}) = nE$, one can evaluate $Z(\beta)$ defined above, as $\sum_E N(E)e^{-\beta E}$ (in the discrete case), which is of the exponential order of $\exp\{n\max_E[S(E) - \beta E]\}$. Defining $\phi(\beta) = \lim_{n\to\infty}[\ln Z(\beta)]/n$, and the free–energy per–particle as $F(\beta) = -\phi(\beta)/\beta$, we obtain the relation $\phi(\beta) = \max_E[S(E) - \beta E]$, where here $E = E(\beta)$ is the maximizer of $[S(E) - \beta E]$. For a given β, the Boltzmann–Gibbs distribution has a sharp peak (for large n) at the level of $E(\beta)$ Joules per–particle. The Legendre relation $\phi(\beta) = \max_E[S(E) - \beta E]$ can be reversed to obtain $S(E) = \min_{\beta\geq 0}[\beta E + \phi(\beta)]$, and both relations can be identified with the thermodynamical definition of the free energy as $F = E - ST$. In the latter relation, the minimizing $\beta = \beta(E)$ (the inverse function of $E(\beta)$) is the equilibrium inverse temperature associated with energy level E. The second law of thermodynamics asserts that in an isolated system (which does not exchange energy with its environment), the total entropy cannot decrease, and hence in equilibrium, it reaches its maximum. When the system is allowed to exchange energy with the environment (in constant volume and temperature), this maximum entropy principle is replaced by the minimum free energy principle: Free energy cannot increase, and it reaches its minimum in equilibrium.

When the Hamiltonian is additive, that is, $\mathcal{E}(\boldsymbol{x}) = \sum_i \mathcal{E}(x_i)$, then $P(\boldsymbol{x})$ has a product form (the particles do not interact), and then the above mentioned physical quantities per particle can be extracted from the case $n = 1$. In this additive case, the Legendre relation $S(E) = \min_{\beta\geq 0}[\beta E + \phi(\beta)]$ is identical to the rate function (the exponent of the Chernoff bound) pertaining to the probability of the event $\sum_{i=1}^n \mathcal{E}(x_i) \leq nE$, thus the parameter to be optimized in the Chernoff bound plays the role of inverse temperature in the corresponding statistical–mechanical system.

Another look at this correspondence between large deviations rate functions and thermal equilibrium is the following: If P is the above mentioned Boltzmann–Gibbs distribution and Q is another probability distribution on the micorstates $\{\boldsymbol{x}\}$, then as is shown e.g., in [2], the Kullback–Leibler divergence between Q and P is given by $D(Q\|P) = \beta(F_Q - F_P)$, where F_P and F_Q are, respectively, the free energies pertaining to P and Q. The rate function pertaining to a large deviations event is normally given by the minimum divergence under the constraints corresponding to this event (see,

[2] Actually, the definition should also include a factor of k, which we will omit in this discussion, thus considering $S(E)$ as the per–particle entropy in units of k.

e.g., [7, Chap. 11]), and so, it is equivalent to minimum free energy, i.e., thermal equilbrium by the second law.

4 Main Result

Let $\mathcal{U}$ and $\mathcal{V}$ be finite[3] sets and let $f : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ be a given function. Let $P = \{p(v),\ v \in \mathcal{V}\}$ be a probability mass function on $\mathcal{V}$ and let $Q = \{q(u|v),\ u \in \mathcal{U},\ v \in \mathcal{V}\}$ be a matrix of conditional probabilities from $\mathcal{V}$ to $\mathcal{U}$. Next, let us define for each $v \in \mathcal{V}$, the (weighted) partition function:

$$Z_v(\beta) = \sum_{u \in \mathcal{U}} q(u|v) e^{-\beta f(u,v)}, \qquad \beta \geq 0, \tag{3}$$

and for a given E_v in the range

$$\min_{u \in \mathcal{U}} f(u,v) \leq E_v \leq \sum_{u \in \mathcal{U}} q(u|v) f(u,v), \tag{4}$$

let

$$S_v(E_v) = \min_{\beta \geq 0} [\beta E_v + \ln Z_v(\beta)]. \tag{5}$$

Further, for a given constant E in the range

$$\sum_{v \in \mathcal{V}} p(v) \min_{u \in \mathcal{U}} f(u,v) \leq E \leq \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} p(v) q(u|v) f(u,v),$$

let

$$\bar{S}(E) = \min_{\beta \geq 0} \left[\beta E + \sum_{v \in \mathcal{V}} p(v) \ln Z_v(\beta) \right]. \tag{6}$$

Let $\mathcal{H}(E)$ denote the set of all $|\mathcal{V}|$–dimensional vectors $\bar{E} = \{E_v,\ v \in \mathcal{V}\}$, where each component E_v satisfies (4), and where $\sum_v p(v) E_v \leq E$. Our main result, in this section, is the following:

Theorem 1

$$\max_{\bar{E} \in \mathcal{H}(E)} \sum_{v \in \mathcal{V}} p(v) S_v(E_v) = \bar{S}(E). \tag{7}$$

[3]The assumption that $\mathcal{U}$ is finite, is made mostly for the sake of convenience and simplicity. Most of our results extend straightforwardly to the case of a continuous alphabet $\mathcal{U}$. The extension to a continuous alphabet $\mathcal{V}$ is somewhat more subtle, however.

The expression on the right–hand side is, of course, more convenient to work with since it involves minimization w.r.t. one parameter only, as opposed to the left–hand side, where there is a minimization over β for every v, as well as a maximization over the $|\mathcal{V}|$–dimensional vector $\bar{E}$.

While the proof of Theorem 1 below is fairly short, in the Appendix (subsection A.1), we outline an alternative proof which, although somewhat longer, provides some additional insight, we believe. As described briefly in the Introduction, it is based on two different approaches to the analysis of the rate function, $I(E)$, pertaining to the probability of the event:

$$\sum_{i=1}^{n} f(U_i, v_i) \leq nE, \tag{8}$$

where $\{U_i\}$ are RV's taking values in $\mathcal{U}$ and drawn according to $q(u^n|v^n) = \prod_{i=1}^{n} q(u_i|v_i)$, and $v^n = (v_1, \ldots, v_n)$ is a given deterministic vector whose components are in $\mathcal{V}$, with each $v \in \mathcal{V}$ appearing n_v times ($\sum_{v\in\mathcal{V}} n_v = n$), and the related relative frequency, n_v/n is exactly $p(v)$.

It should be noted that the proof in the Appendix pertains to a slightly different definition of the set $\mathcal{H}(E)$, where the individual upper bound to each E_v is enlarged to $\max_u f(u, v)$. Thus, $\mathcal{H}(E)$ is extended to a larger set, which will be denoted by $\mathcal{H}_0(E)$ in the Appendix. But the maximum over $\mathcal{H}_0(E)$ is always attained within the original set $\mathcal{H}(E)$ (as is actually shown in the proof below).

Proof. Here we prove the identity of Theorem 1 directly, without using large deviations analysis and Chernoff bounds. We first prove that for every $\bar{E} \in \mathcal{H}(E)$, we have $\sum_{v\in\mathcal{V}} p(v)S_v(E_v) \leq \bar{S}(E)$ and then, of course,

$$\max_{\bar{E}\in\mathcal{H}(E)} \sum_{v\in\mathcal{V}} p(v)S_v(E_v) \leq \bar{S}(E)$$

as well. This follows from the following chain of inequalities:

$$\begin{aligned}
\sum_{v\in\mathcal{V}} p(v)S_v(E_v) &= \sum_{v\in\mathcal{V}} p(v) \cdot \min_{\beta\geq 0}[\beta E_v + \ln Z_v(\beta)] \\
&= \sum_{v\in\mathcal{V}} \min_{\beta\geq 0}[\beta p(v) E_v + p(v) \ln Z_v(\beta)] \\
&\leq \min_{\beta\geq 0}\left[\beta \sum_{v\in\mathcal{V}} p(v)E_v + \sum_{v\in\mathcal{V}} p(v) \ln Z_v(\beta)\right] \\
&\leq \min_{\beta\geq 0}\left[\beta E + \sum_{v\in\mathcal{V}} p(v) \ln Z_v(\beta)\right] \\
&= \bar{S}(E),
\end{aligned} \tag{9}$$

where in the second inequality we used the postulate that $\sum_v p(v)E_v \leq E$.

In the other direction, let β^* be the achiever of $\bar{S}(E)$, i.e., β^* is the solution to the equation:

$$E = -\left[\frac{\partial}{\partial\beta}\sum_v p(v)\ln Z_v(\beta)\right]_{\beta=\beta^*}.$$

For each $v \in \mathcal{V}$, let $E_v^* \in [\min_u f(u,v), \sum_u q(u|v)f(u,v)]$ be chosen such that β^* would be the achiever of $S_v(E_v^*)$, i.e., $E_v^* = -[\partial \ln Z_v(\beta)/\partial\beta]_{\beta=\beta^*}$. Obviously, the vector $\{E_v^*,\ v \in \mathcal{V}\}$ lies in $\mathcal{H}(E)$, and

$$\begin{aligned}\sum_v p(v)E_v^* &= -\sum_v p(v)\left[\frac{\partial \ln Z_v(\beta)}{\partial\beta}\right]_{\beta=\beta^*} \\ &= -\left[\frac{\partial}{\partial\beta}\sum_v p(v)\ln Z_v(\beta)\right]_{\beta=\beta^*} \\ &= E. \end{aligned} \tag{10}$$

Thus,

$$\begin{aligned}\max_{\bar{E}\in\mathcal{H}(E)}\sum_{v\in\mathcal{V}} p(v)S_v(E_v) &\geq \sum_{v\in\mathcal{V}} p(v)S_v(E_v^*) \\ &= \sum_{v\in\mathcal{V}} p(v)[\beta^* E_v^* + \ln Z_v(\beta^*)] \\ &= \beta^*\sum_{v\in\mathcal{V}} p(v)E_v^* + \sum_v p(v)\ln Z_v(\beta^*) \\ &= \beta^* E + \sum_v p(v)\ln Z_v(\beta^*) \\ &= \bar{S}(E). \end{aligned} \tag{11}$$

This completes the proof of Theorem 1. □

The function $Z_v(\beta)$ is similar to the well–known partition function pertaining to the Boltzmann distribution w.r.t. the Hamiltonian $\mathcal{E}_v(u) = f(u,v)$, except that each exponential term is weighted by $q(u|v)$, as opposed to the usual form, which is just $\sum_{u\in\mathcal{U}} e^{-\beta\mathcal{E}_v(u)}$. Before describing the statistical mechanical interpretation of eq. (7), we should note that $Z_v(\beta)$ defined in (3) can easily be related to the ordinary partition function, without weighting, as follows: Suppose that $\{q(u|v)\}$ are rational[4] and hence can be represented as ratios of two positive integers, $q(u|v) = M(u|v)/M$, where $M \geq |\mathcal{U}|$ is common to all $u \in \mathcal{U}$ (and $v \in \mathcal{V}$). Now, imagine that every value of u actually

[4]Even if not rational, they can always be approximated as such to an arbitrarily good precision.

represents a 'quantization' of a more refined microstate (call it a "nanostate") $w \in \mathcal{W}$, $|\mathcal{W}| = M$, so that $u = g_v(w)$, where g_v is a many–to–one function, for which the inverse image of every u consists of $M(u|v)$ many values of w. Suppose further that the Hamiltonian depends on w only via $g_v(w)$, i.e., $\mathcal{E}'_v(w) = \mathcal{E}_v(g_v(w))$. Then, the (ordinary) partition function related to w is given by

$$\begin{aligned}
\zeta_v(\beta) &= \sum_{w\in\mathcal{W}} e^{-\beta\mathcal{E}'_v(w)} \\
&= \sum_{w\in\mathcal{W}} e^{-\beta\mathcal{E}_v(g_v(w))} \\
&= \sum_{u\in\mathcal{U}} M(u|v)e^{-\beta\mathcal{E}_v(u)} \\
&= M\sum_{u\in\mathcal{U}} q(u|v)e^{-\beta\mathcal{E}_v(u)} = MZ_v(\beta). \qquad (12)
\end{aligned}$$

Thus, the weighted partition function is, within a constant factor M, the same as the ordinary partition function of w. This factor cancels out when probabilities are calculated since it appears both in the numerator and the denominator. Moreover, it affects neither the minimizing β that achieves $S_v(E_v)$ or $\bar{S}(E)$, nor the derivatives of the log–partition function.

We now move on to our interpretation of eq. (7) from the viewpoint of elementary statistical physics: Consider a physical system which consists of $|\mathcal{V}|$ subsystems of particles. The total number of particles in the system is n and the total amount of energy is nE. For each $v \in \mathcal{V}$, the subsystem indexed by v (subsystem v, for short) contains $n_v = np(v)$ particles, each of which can lie in any microstate within a finite set of microstates $\mathcal{U}$ (or an underlying nanostate in a set $\mathcal{W}$), and it is characterized by an additive Hamiltonian $\sum_{i:\ v_i=v} \mathcal{E}_v(u_i)$, where $\mathcal{E}_v(u) = f(u, v)$. The total amount of energy possessed by subsystem v is given by n_vE_v. As long as the subsystems are in thermal isolation from each other, each one of them may have its own temperature $T_v = 1/(k\beta_v)$, where β_v is the achiever of the normalized (per–particle) entropy associated with an average per–particle energy E_v, i.e.,

$$S_v(E_v) = \min_{\beta\geq 0}[\beta E_v + \ln Z_v(\beta)].$$

The above–mentioned rate function $I(E)$ of $\Pr\{\sum_{i=1}^n f(U_i, v_i) \leq nE\}$ is then given by the negative maximum total per–particle entropy, $\sum_v p(v)S_v(E_v)$, where the maximum is over all energy allocations $\{E_v\}$ such that the total energy is conserved, i.e., $\sum_v p(v)E_v = E$. This maximum is attained by the expression of the r.h.s. of eq. (7), where there is *only one* temperature parameter,

and hence it corresponds to *thermal equilibrium*. In other words, the whole system then lies in the same temperature $T^* = 1/(k\beta^*)$, where β^* is the minimizer of $\bar{S}(E)$. Thus, the energy allocation among the various subsystems (pertaining to the dominant configuration associated with the large deviations event under discussion) is in thermal equilibrium – their temperatures are the same (cf. the above proof of Theorem 1). Theorem 1 is then interpreted as expressing the second law of thermodynamics.

At this point, a few comments are in order:

1. It should be pointed out that in the above physical interpretation, we have implicitly assumed that the particles within each subsystem are distinguishable, and so the partition function corresponding to a set of n_v particles is given by the partition function of a single particle raised to the power of n_v.

2. As mentioned in the above paragraph, our conclusion is that $I(E) = -\bar{S}(E)$. One should keep in mind that $\bar{S}(E)$ is induced by a (convex) combination of weighted partition functions, rather than ordinary partition functions, like $\zeta_v(\beta)$. Referring to eq. (12), the ordinary notion of entropy $\Sigma(E)$ as the normalized log–number of (nano)states with normalized energy E, is given by

$$\begin{aligned}\bar{\Sigma}(E) &= \min_{\beta\geq 0}\left[\beta E + \sum_v p(v)\ln\zeta_v(\beta)\right] \\ &= \min_{\beta\geq 0}\left[\beta E + \sum_v p(v)\ln Z_v(\beta)\right] + \ln M \\ &= \bar{S}(E) + \ln M. \end{aligned} \tag{13}$$

   Thus,

$$I(E) = \ln M - \bar{\Sigma}(E),$$

   which is always non–negative.

3. The identity (7) can be thought of as a generalized concavity property of the entropy: Had all the entropy functions $S_v(\cdot)$ been the same, this would have been the usual concavity property. What makes this equality less trivial and more interesting is that it continues to hold even when $S_v(\cdot)$, for the various $v \in \mathcal{V}$, are different from each other.

4. On the more technical level, since this paper draws analogies with physics, we should say a few words about physical units. The products βE, βE_v, $\beta f(u,v)$, etc., should all be pure numbers, of course. Since $\beta = 1/(kT)$, where k is Boltzmann's constant and T is absolute temperature, and since kT has units of energy, it is understood that E, E_v, $f(u,v)$ and the like, should all have units of energy as well. In the applications described below, whenever this is not the case, i.e., the latter quantities are pure numbers rather than physical energies, we will sometimes reparametrize β by $\beta\epsilon_0$, where ϵ_0 is an arbitrary constant possessing units of energy (e.g., $\epsilon_0 = 1$ Joule), and we absorb ϵ_0 in the Hamiltonian, i.e., redefine $\mathcal{E}_v(u) = \epsilon_0 f(u,v)$. Thus, in this case, $S_v(E)$, where E is the now the energy in units of ϵ_0, is redefined as

$$S_v(E) = \min_{\beta \geq 0} \left[\beta \cdot \epsilon_0 E + \ln \left(\sum_u q(u|v) e^{-\beta \mathcal{E}_v(u)} \right) \right].$$

This kind of modification is not essential, but it may help to avoid confusion about units when the picture is viewed from the aspects of physics.

5 Applications

Equipped with the main result of the previous section and its statistical mechanical interpretation, we next introduce a few applications that fall within the framework considered. In all these applications, there is an underlying large deviations event of the type of eq. (8), whose rate function is of interest. None of the information–theoretic results here are new, but the above described viewpoint of statistical physics is relevant to all of them, and it sheds some new light on these application examples, we believe.

5.1 The Rate–Distortion Function

Let $P = \{p(x),\ x \in \mathcal{X}\}$ designate the vector of letter probabilities associated with a given discrete memoryless source (DMS), and for a given reproduction alphabet $\hat{\mathcal{X}}$, let $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ denote a single–letter distortion measure. Let $R(D)$ denote the rate–distortion function of the DMS P.

One useful way to think of the rate–distortion function is inspired by the classical random coding argument: Let $(\hat{X}_1, \ldots, \hat{X}_n)$ be drawn i.i.d. from the optimum random coding distribution $q^*(\hat{x}_1, \ldots, \hat{x}_n) = \prod_{i=1}^n q^*(\hat{x}_i)$ and consider the event $\sum_{i=1}^n d(x_i, \hat{X}_i) \leq nD$, where x^n is a given source

vector, typical to P, i.e., the composition of x^n consists of $n_x = np(x)$ occurrences of each $x \in \mathcal{X}$. This is exactly an event of the type (8) with $U_i = \hat{X}_i$, $v_i = x_i$, $i = 1, \ldots, n$, $q(u|v) = q(\hat{x}|x) = q^*(\hat{x})$ independently of x, $f(u, v) = f(\hat{x}, x) = d(x, \hat{x})$, and $E = D$. I.e., the Hamiltonian $\mathcal{E}_x(\hat{x})$ is given by $\epsilon_0 d(x, \hat{x})$ and the total energy is nD in units of ϵ_0.

Suppose that this probability is of the exponential order of $e^{-nI(D)}$. Then, it takes about $M = e^{n[I(D)+\epsilon]}$ ($\epsilon > 0$, however small) independent trials to 'succeed' at least once (with high probability) in having some realization of $\hat{X}^n$ within distance nD from x^n. This is the well–known classical random coding achievability argument that leads to $I(D) = R(D)$. Thus, the large–deviations rate function of interest agrees exactly with the rate–distortion function (cf. [4, Sect. 3.4]), which is:

$$R(D) = -\min_{\beta \geq 0} \left[\beta \cdot \epsilon_0 D + \sum_{x \in \mathcal{X}} p(x) \ln \left(\sum_{\hat{x} \in \hat{\mathcal{X}}} q^*(\hat{x}) e^{-\beta \cdot \epsilon_0 d(x, \hat{x})} \right) \right]. \tag{14}$$

Indeed, we thus recover the formula of [12, p. 90, Corollary 4.2.3]) for the rate–distortion function, which is shown there, using completely different considerations. Nonetheless, this random coding point of view applies to any (i.i.d.) random coding distribution (not only the optimum one, that attains $R(D)$), and our discussion still applies.

The fact that the rate–distortion function has an interpretation of an isothermal equilibrium situation in statistical thermodynamics is not quite new (cf. e.g. [4, Sect. 6.4], [6],[44]). Here, however, we obtain it in a more explicit manner and as a special case of a more general principle.

A simple example is that of the binary symmetric source with the Hamming distortion measure. It is easy to see that, in this example, the relationship between distortion and temperature is:

$$T = \frac{\epsilon_0}{k \ln[(1-D)/D]} \quad \text{or, equivalently,} \quad D = \frac{1}{1 + e^{\epsilon_0/(kT)}} \tag{15}$$

and, of course, $R(D) = 1 - h_2(D)$, where $h_2(D)$ is the binary entropy function.

A slightly more involved example pertains to the regime of high resolution (small distortion) of continuous–alphabet sources and it turns out to be related to (a generalized version of) the law of equipartition of energy in statistical physics: Consider the L_θ distortion measure, $d(x, \hat{x}) = |x - \hat{x}|^\theta$ (most commonly encountered are the cases $\theta = 1$ and $\theta = 2$). Let us assume that $D > 0$ is very small and consider the (continuous) uniform random coding distribution $q(\hat{x}) = \frac{1}{2A}$ in the interval

$[-A, A]$ and zero elsewhere. This random coding distribution is suboptimal, but it corresponds, and hence is well motivated, by many results in high–resolution quantization using uniform quantizers (see, e.g., [14] and references therein). For every $x \in \mathcal{X}$, the partition function is given by

$$Z_x(\beta) = \frac{1}{2A} \int_{-A}^{A} \exp\{-\beta\epsilon_0 |\hat{x} - x|^\theta\} \mathrm{d}\hat{x}.$$

The average distortion (internal energy) associated with this partition function can be derived using essentially the same technique as the one that leads to the law of equipartition in statistical physics (although somewhat more care should be exercised since $Z_x(\beta)$ depends on x):

$$\begin{aligned}
\epsilon_0 D &= -\frac{\partial}{\partial\beta} \ln \left[\int_{-A}^{A} \exp\{-\beta\epsilon_0 |\hat{x} - x|^\theta\} \mathrm{d}\hat{x} \right] \\
&= -\frac{\partial}{\partial\beta} \ln \left[\beta^{-1/\theta} \cdot \int_{-\beta^{1/\theta}(A+x)}^{\beta^{1/\theta}(A-x)} \exp\{-\epsilon_0 |\beta^{1/\theta}(\hat{x} - x)|^\theta\} \mathrm{d}(\beta^{1/\theta}(\hat{x} - x)) \right] \\
&= -\frac{\partial}{\partial\beta} \ln \left[\beta^{-1/\theta} \cdot \int_{-\beta^{1/\theta}(A+x)}^{\beta^{1/\theta}(A-x)} \exp\{-\epsilon_0 |z|^\theta\} \mathrm{d}z \right] \\
&= -\frac{\partial}{\partial\beta} \ln \left(\beta^{-1/\theta} \right) - \frac{\partial}{\partial\beta} \ln \left[\int_{-\beta^{1/\theta}(A+x)}^{\beta^{1/\theta}(A-x)} \exp\{-\epsilon_0 |z|^\theta\} \mathrm{d}z \right] \\
&= \frac{1}{\beta\theta} \left\{ 1 - \frac{\beta^{1/\theta}[(A-x)\exp\{-\beta\epsilon_0 |A-x|^\theta\} + (A+x)\exp\{-\beta\epsilon_0 |A+x|^\theta\}]}{\int_{-\beta^{1/\theta}(A+x)}^{\beta^{1/\theta}(A-x)} \exp\{-\epsilon_0 |z|^\theta\} \mathrm{d}z} \right\}. \quad (16)
\end{aligned}$$

When β is very large, the denominator of the second term of the expression in the curly brackets of the right–most side, goes to $\int_{-\infty}^{\infty} \exp\{-\epsilon_0 |z|^\theta\} \mathrm{d}z$, which is a constant. Now if, in addition, $|x| < A$, then the numerator tends to zero as β grows without bound. Thus, the dominant term, for low temperatures, is $1/(\beta\theta) = kT/\theta$. Note that for $\theta = 2$, where the Hamiltonian is quadratic in the integration variable $\hat{x}$, we obtain $kT/2$, which is exactly the law of equipartition.[5] An exact closed–form expression, for every finite β, can be derived for the case $\theta = 1$, since in this case, the integral at the denominator has a simple expression. For example, setting $\theta = 1$, and $x = 0$ in the above expression, yields:

$$\begin{aligned}
D &= \frac{1}{\beta\epsilon_0} - \frac{A}{e^{\beta\epsilon_0 A} - 1} \\
&= \frac{kT}{\epsilon_0} - \frac{A}{e^{\epsilon_0 A/(kT)} - 1}. \quad (17)
\end{aligned}$$

[5]The physical analogue of this case is the Hamiltonian pertaining to an harmonic oscillator whose equilibrium point is x.

It should be pointed out that this expression is valid only in the range where it is monotonically increasing in T. (Beyond this point, the minimizing β is no longer the point of zero derivative).

Returning to the case of a general θ, it should be noted that the relation that we have just shown, $D \approx kT/(\epsilon_0\theta)$ (for low temperatures), can also be obtained as a special case of a further generalization of the equipartition theorem (see, e.g., [26, p. 61, Exercise 13.6]), asserting (in our notation) that if $\hat{x} \to \infty$ implies $\mathcal{E}(\hat{x}) \to \infty$, then the mean of $\hat{x} \cdot \mathcal{E}'(\hat{x})$ is kT, where $\mathcal{E}'$ is the derivative of $\mathcal{E}$.

Since $D \approx kT/(\epsilon_0\theta)$ for small T, as mentioned, i.e., the distortion (energy) is linear in temperature in that regime, the constant of proportionality must be related to the heat capacity per particle C (a.k.a. specific heat) which is the dervative of the energy w.r.t. temperature, that is, $C = k/\theta$. Since the temperature is proportional to the negative local slope of the distortion–rate function (as the reciprocal, β, is proportional to the negative local slope of the rate–distortion function), this means that the distortion is proportional to its derivative w.r.t. R, which means an exponential relationship of the form $D = D_0 e^{-\theta R}$ (D_0 – constant). For $\theta = 2$ (mean square error), this is recognized as the well–known characterization of distortion as function of rate in the high resolution regime. Specifically, in this case, the factor of 2 at the denominator of $kT/2$, the universal expression of the internal energy per degree of freedom according to the equipartition theorem, has the same origin as the factor of 2 that appears in the exponent of $D(R) = D_0 e^{-2R}$ (decay of 6dB per bit). Thus, the law of equipartition in statistical physics is related to the behavior of rate distortion codes in the high resolution regime.

To compute the rate associated with this temperature more explicitly, note that the minimizing β^* is given by $1/(\theta\epsilon_0 D)$, and so

$$\begin{aligned} R &= -\beta^*\epsilon_0 D - \ln\left[\frac{1}{2A}\int_{-\infty}^{\infty}\exp\{-\beta^*\epsilon_0|\hat{x}-x|^\theta\}\mathrm{d}\hat{x}\right] \\ &= -\frac{1}{\theta} - \ln\left[\frac{1}{2A}\cdot\frac{2\Gamma(1/\theta)}{\theta(1/\theta D)^{1/\theta}}\right] \\ &= \ln\left[\frac{A\theta}{\Gamma(1/\theta)(\theta e D)^{1/\theta}}\right] \\ &= \ln\left[\frac{A\theta}{\Gamma(1/\theta)}\right] - \frac{1}{\theta}\ln(\theta e D). \end{aligned} \tag{18}$$

5.2 Channel Capacity

In complete duality to the random coding argument that puts the rate–distortion function in the framework discussed in Section 3, a parallel argument can be made with regard to channel capacity, and again, it can be extended to any i.i.d. random coding distribution, not only the capacity–achieving distribution (cf. also [6]).

Given a discrete memoryless channel (DMC) with a finite input alphabet $\mathcal{X}$, and a finite output alphabet $\mathcal{Y}$, we can obtain capacity using the following argument. Let $\{q^*(x),\ x \in \mathcal{X}\}$ be the optimum random coding distribution according to which, each codeword X^n is drawn independently. Let y^n be a given channel output sequence which is typical to the output distribution $p(y) = \sum_{x\in\mathcal{X}} q(x)W(y|x)$, where $\{W(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ are the channel transition probabilities. That is, each symbol y appears $n_y \sim np(y)$ times in y^n. Consider now a threshold decoder that decides on a message if its corresponding likelihood score $\sum_{i=1}^n \log \frac{1}{W(y_i|X_i)}$ is uniquely below the threshold $nH(Y|X)$, where $H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} q(x)W(y|x)\log W(y|x)$, and declares an error otherwise (i.e., if none or more than one message meets this criterion). Consider then the large deviations event

$$\sum_{i=1}^{n} \log \frac{1}{W(y_i|X_i)} \leq nH(Y|X). \tag{19}$$

By the union bound, as long as the number of randomly chosen codewords is exponentially less than e^{-nI}, where I is the rate function of the large–deviations event (19), then the average error probability still vanishes as $n \to \infty$. Since this is the exactly the achievability argument of the channel coding theorem, then $I = C$, where C the channel capacity.

Again, this complies with our model setting with the assignments, $U_i = X_i$, $v_i = y_i$, $i = 1, \ldots, n$, $q(u|v) = q(x|y) = q^*(x)$ independently of y, $f(u, v) = f(x, y) = -\log W(y|x)$ and $E = H(Y|X)$ units of ϵ_0. In other words, channel capacity can be represented as

$$C = -\min_{\beta\geq 0}\left[\beta\cdot\epsilon_0 H(Y|X) + \sum_{y\in\mathcal{Y}} p(y)\ln\left(\sum_{x\in\mathcal{X}} q^*(x)e^{-\beta\cdot\epsilon_0[-\log W(y|x)]}\right)\right]. \tag{20}$$

It is easy to see that, in this case, the equilibrium temperature always corresponds to $\beta\epsilon_0 = 1$, namely, $T = \epsilon_0/k$ (cf. the Nishimori temperature [17], [38]).

By the same token, one can derive an expression of the random coding capacity pertaining to mismatched decoding, where the decoder uses an additive metric $m(x, y)$ other than the optimum

metric, $-\log W(y|x)$ (see, e.g., [3], [9], [28], [29], [32], and references therein). The only modifications to the above expression would be to replace the Hamiltonian by $\mathcal{E}_y(x) = \epsilon_0 m(x,y)$ and to replace $H(Y|X)$ by the expectation of $m(X,Y)$ w.r.t. $q^*(x)W(y|x)$. The new optimum random coding distribution might change as well. Here, it is no longer necessarily true that the equilibrium temperature is $T = \epsilon_0/k$.

5.3 Signal Detection and Hypothesis Testing

Consider the following binary hypothesis testing problem: Given a deterministic signal, which is represented by a sequence $x^n = (x_1, \ldots, x_n)$ with elements taking on values in a (finite) set $\mathcal{X}$ and relative frequencies $\{p(x),\ x \in \mathcal{X}\}$, and given an observation sequence $Y^n = (Y_1, \ldots, Y_n)$, we are required to decide between two hypotheses:

H_0 : The observation vector Y^n is "pure noise," distributed according to some product measure $Q = \{q(y),\ y \in \mathcal{Y}\}$, i.e., $q(y^n) = \prod_{i=1}^n q(y_i)$, which is unrelated to x^n.

H_1 : The observation vector Y^n is a "noisy version" of x^n, distributed according to $q(y^n|x^n) = \prod_{i=1}^n q(y_i|x_i)$.

The optimum detector (under both the Bayesian and the Neyman–Pearson criterion) compares the likelihood ratio $\sum_{i=1}^n \ln[q(y_i)/q(y_i|x_i)]$ to a threshold nE_0, and decides in favor of H_0 if this threshold is exceeded, otherwise, it decides in favor of H_1.

The false–alarm probability then is the probability of the event

$$\sum_{i=1}^{n} \ln\left[\frac{q(Y_i)}{q(Y_i|x_i)}\right] \leq nE_0$$

under Q. This, again, fits our scenario with the substitutions $U_i = Y_i$, $v_i = x_i$, $i = 1, \ldots, n$, $q(u|v) = q(y)$, independently of $x = v$, $f(u,v) = f(y,x) = \ln[q(y)/q(y|x)]$, and $E = E_0$. Similarly, the analysis of the missed–detection probability corresponds to the assignments: $U_i = Y_i$, and $v_i = x_i$, $i = 1, \ldots, n$, as before, but now $q(u|v) = q(y|x)$, $f(u,v) = f(y,x) = \ln[q(y|x)/q(y)]$ and $E = -E_0$. Note that when $\{q(y)\}$ is the uniform distribution over $\mathcal{Y}$, the missed-detection event can also be interpreted as the probability of excess code–length of an arithmetic lossless source code w.r.t. $\{q(y|x)\}$.

Another situation of hypothesis testing that is related to our study in a similar manner is one where the signal x^n is always underlying the observations, but the decision to be made is associated with two hypotheses regarding the noise level, or the temperature. In this case, there is a certain Hamiltonian $\mathcal{E}_x(y)$ for each $x \in \mathcal{X}$, and we assume a Boltzmann–Gibbs distribution parametrized by the temperature

$$q(y|x,\beta) = \frac{e^{-\beta\mathcal{E}_x(y)}}{\zeta_x(\beta)}$$

where

$$\zeta_x(\beta) = \sum_y e^{-\beta\mathcal{E}_x(y)}.$$

Note that here $\zeta_x(\beta)$ is an ordinary partition function, without weighting (cf. (12)). We shall also denote

$$\bar{\Sigma}(E) = \min_{\beta\geq 0}\left[\beta E + \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta)\right].$$

As $\bar{\Sigma}(E)$ is induced by a convex combination of non-weighted partition functions, it has the significance of the normalized logarithm of the number of microstates with energy about nE. Thus, $\bar{\Sigma}(E)$ s the thermodynamic entropy.

Given two values β_1 and β_2 (say, $\beta_1 > \beta_2$), the hypotheses now are the following:

$H_1 :$ Y^n is distributed according to $q_1(y^n|x^n) = \prod_{i=1}^n q(y_i|x_i,\beta_1)$.

$H_2 :$ Y^n is distributed according to $q_2(y^n|x^n) = \prod_{i=1}^n q(y_i|x_i,\beta_2)$.

The likelihood ratio test compares $\sum_{i=1}^n \mathcal{E}_{x_i}(Y_i)$ to a threshold, nE_0, and decides in favor of H_2 if the threshold is exceeded, otherwise, it favors H_1. Here, E_0 should lie in the interval (E_1, E_2), where

$$E_i \stackrel{\Delta}{=} -\sum_{x\in\mathcal{X}} p(x)\cdot\left[\frac{\partial\ln\zeta_x(\beta)}{\partial\beta}\right]_{\beta=\beta_i}, \quad i=1,2.$$

For convenience, let us assume now that E_i, $i=0,1,2$, and $\mathcal{E}_x(y)$ already have units of energy, so there is no need to have the constant ϵ_0. In this situation, the exponent of the error probability

under H_2 is given by $-\bar{S}(E_0)$, where

$$\begin{aligned}
\bar{S}(E_0) &= \min_{\beta\geq 0}\left[\beta E_0 + \sum_{x\in\mathcal{X}} p(x)\ln\left(\sum_{y\in\mathcal{Y}} q(y|x,\beta_2)e^{-\beta\mathcal{E}_x(y)}\right)\right] \\
&= \min_{\beta\geq 0}\left[\beta E_0 + \sum_{x\in\mathcal{X}} p(x)\ln\left(\frac{\zeta_x(\beta+\beta_2)}{\zeta_x(\beta_2)}\right)\right] \\
&= \min_{\beta\geq 0}\left[\beta E_0 + \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta+\beta_2) - \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta_2)\right] \\
&= \min_{\beta\geq 0}\left[(\beta+\beta_2)E_0 + \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta+\beta_2)\right] - \beta_2 E_0 - \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta_2) \\
&= \min_{\beta\geq\beta_2}\left[\beta E_0 + \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta)\right] + \beta_2(E_2-E_0) - \left[\beta_2 E_2 + \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta_2)\right] \\
&= \min_{\beta\geq\beta_2}\left[\beta E_0 + \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta)\right] + \beta_2(E_2-E_0) \\
&\quad - \min_{\beta\geq 0}\left[\beta E_2 + \sum_{x\in\mathcal{X}} p(x)\ln\zeta_x(\beta)\right] \\
&= \bar{\Sigma}(E_0) - \bar{\Sigma}(E_2) + \beta_2(E_2-E_0), \qquad (21)
\end{aligned}$$

where we have used the fact that the achiever $\beta(E)$ of $\bar{\Sigma}(E)$ is a monotonically non-increasing function of E, thus, $E_0 < E_2$ implies $\beta(E_0) > \beta(E_2) = \beta_2$, and so, the global minimum over $\beta \geq 0$ is attained for $\beta \geq \beta_2$ anyway.

It then follows that the error exponent I_2 under H_2 is given by

$$\begin{aligned}
I_2 &= \bar{\Sigma}(E_2) - \bar{\Sigma}(E_0) - \beta_2(E_2-E_0) \\
&= \frac{1}{k}\left[k\bar{\Sigma}(E_2) - k\bar{\Sigma}(E_0) - \frac{E_2-E_0}{T_2}\right] \\
&= \frac{1}{k}\int_{E_0}^{E_2}\left[\frac{1}{T(E)} - \frac{1}{T_2}\right]\mathrm{d}E \\
&= \frac{1}{k}\int_{T_0}^{T_2}\left(\frac{1}{T} - \frac{1}{T_2}\right)\bar{C}(T)\mathrm{d}T, \qquad (22)
\end{aligned}$$

where $T(E) = 1/(k\beta(E))$ is the temperature corresponding to energy E, $T_i = T(E_i)$, $i = 0, 1, 2$, and $\bar{C}(T) = \mathrm{d}E/\mathrm{d}T$ is the average heat capacity per particle of the system, which is the weighted average of heat capacities of all subsystems, i.e.,

$$\bar{C}(T) = \sum_{x\in\mathcal{X}} p(x)C_x(T),$$

where

$$C_x(T) = \frac{\mathrm{d}E_x}{\mathrm{d}T} = \frac{1}{kT^2}\left[\frac{\mathrm{d}^2 \ln \zeta_x(\beta)}{d\beta^2}\right]_{\beta=1/(kT)}.$$

Thus,

$$I_2 = \sum_{x\in\mathcal{X}} p(x) \cdot \frac{1}{k}\int_{T_0}^{T_2}\left(\frac{1}{T}-\frac{1}{T_2}\right) C_x(T)\mathrm{d}T,$$

which is interpreted as the weighted average of the relative contributions of all subsystems, which all lie in the same temperature T_0.

In a similar manner, the rate function I_1 of the probability of error under H_1 is given by:

$$\begin{aligned} I_1 &= \bar{\Sigma}(E_1) - \bar{\Sigma}(E_0) - \beta_1(E_1 - E_0) \\ &= \frac{1}{k}\left[k\bar{\Sigma}(E_1) - k\bar{\Sigma}(E_0) - \frac{E_1 - E_0}{T_1}\right] \\ &= \frac{1}{k}\int_{E_1}^{E_0}\left[\frac{1}{T_1} - \frac{1}{T(E)}\right]\mathrm{d}E \\ &= \frac{1}{k}\int_{T_1}^{T_0}\left(\frac{1}{T_1}-\frac{1}{T}\right)\bar{C}(T)\mathrm{d}T. \end{aligned} \tag{23}$$

The expression in the square brackets of the second line pertaining to I_2 has a simple graphical interpretation (see Fig. 1): It is the vertical distance (corresponding to the vertical line $E = E_0$) between the curve $\bar{\Sigma}(E)$ and the line tangent to that curve at $E = E_2$ (whose slope is $\beta_2 = \beta(E_2)$). The two other expressions of I_2, in the last chain of equalities, describe the error exponent I_2 in terms of slow heating from temperature T_0 to temperature T_2. Similar comments apply to I_1 (cf. Fig. 1). Thus, the error exponents are linear functionals of the average heat capacity, $\bar{C}(T)$, in the range of temperatures $[T_1, T_2]$. The higher is the heat capacity, the better is the discrimination between the hypotheses. This is related to the fact that Fisher information of the parameter β is given by

$$J(\beta) = \sum_{x\in\mathcal{X}} p(x)\frac{\mathrm{d}^2 \ln \zeta_x(\beta)}{\mathrm{d}\beta^2} = kT^2\bar{C}(T),$$

namely, again, a linear function of $\bar{C}(T)$. However, while the Fisher information depends only on one local value of $\bar{C}(T)$ (as it measures the sensitivity of the likelihood function to the parameter in a local manner), the error exponents depend on $\{\bar{C}(T) : T_1 \le T \le T_2\}$ in a cumulative manner, via the above integrals. The tradeoff between I_1 and I_2 is also obvious: by enlarging the threshold E_0, or, correspondingly, T_0, the range of integration pertaining to I_1 increases at the expense of

the one of I_2 and vice versa. In the extreme case, where $I_2 = 0$, we get

$$I_1 = D(P_2 \| P_1) = \frac{1}{k} \int_{T_1}^{T_2} \left(\frac{1}{T_1} - \frac{1}{T} \right) \bar{C}(T) \mathrm{d}T.$$

These relationships between error exponents, divergences, Fisher information, and heat capacity are new to the best knowledge of the author.

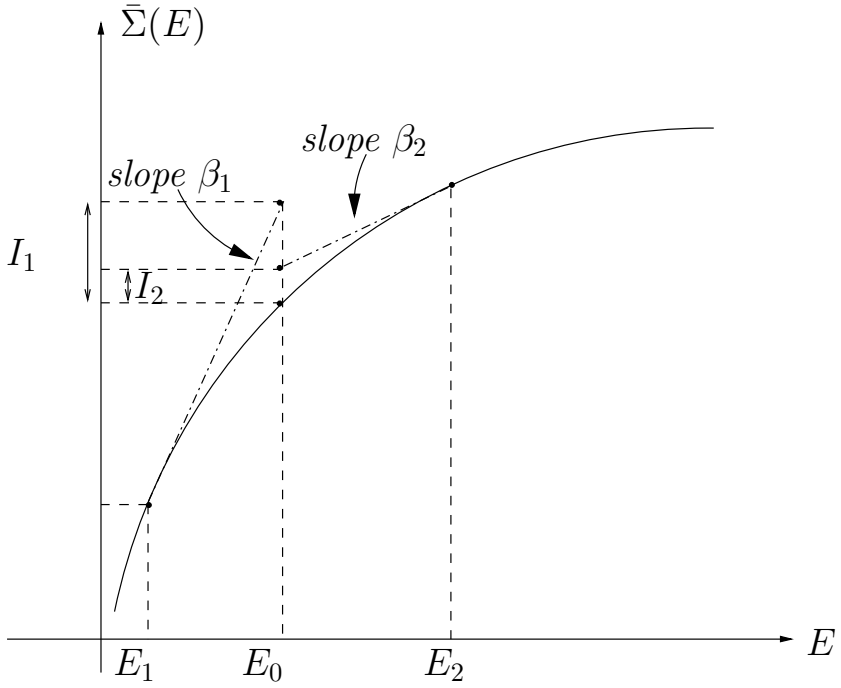

Figure 1: Entropy as function of energy and a graphical representation of error exponents.

5.4 Error Exponents of Time–Varying Scalar Quantizers

In this application example, we are back to the problem area of lossy data compression, but this time, it is about scalar (symbol–by–symbol) compression. This setup is motivated by earlier results about the optimality of time–shared scalar quantizers within the class of causal source codes for memoryless sources, both under the average rate/distortion criteria [37] and large–deviations performance criteria [33]. In particular, it was shown that under both criteria, optimum time–sharing between at most two (entropy coded) scalar quantizers is as good as any causal source code for memoryless sources. Here, we will focus on the large deviations performance criteria, namely, source coding exponents.

Consider a time–varying scalar quantizer $\hat{X}_i = f_i(X_i)$, acting on a DMS $X_1, X_2, \ldots$, $X_i \in \mathcal{X}$, drawn from q, where $\{f_i\}$ is an arbitrary (deterministic) sequence of quantizers from a given finite set $\mathcal{F} = \{F_1, \ldots, F_S\}$, where $F_s : \mathcal{X} \to \hat{\mathcal{X}}_s$, $\hat{\mathcal{X}}_s$ being the reproduction alphabet corresponding to F_s, $s = 1, \ldots, S$. In other words, for every $i = 1, 2, \ldots, n$, $f_i = F_{s_i}$, for a certain arbitrary sequence of 'states', $s_1, s_2, \ldots$ (known to the decoder) with components in $\mathcal{S} = \{1, 2, \ldots, S\}$.

The distortion incurred by such a time–varying scalar quantizer, over n units of time, is $\sum_{i=1}^n d(X_i, f_i(X_i)) = \sum_{i=1}^n d(X_i, F_{s_i}(X_i))$. The total code length is $\sum_{i=1}^n L_{s_i}(F_{s_i}(X_i))$, where the per–symbol length functions $L_s(\cdot)$ may correspond to either fixed–rate coding, where $L_s(\hat{x}) = R_s \stackrel{\Delta}{=} \lceil \log |\hat{\mathcal{X}}_s| \rceil$ for all $\hat{x}$, or any other length function satisfying the Kraft inequality, $\sum_{\hat{x} \in \hat{\mathcal{X}}_s} 2^{-L_s(\hat{x})} \leq 1$. For the sake of simplicity of the exposition, let us assume fixed–rate coding. We will denote by n_s, $s \in \mathcal{S}$, the number of times that $s_i = s$ occurs in s^n, and $p(s) = n_s/n$ is the corresponding relative frequency.

In [33], among other results, the rate function of the excess distortion event

$$\sum_{i=1}^n d(X_i, F_{s_i}(X_i)) > nD, \qquad D > \sum_{(x,s) \in \mathcal{X} \times \mathcal{S}} q(x)p(s)d(x, F_s(x))$$

was optimized across the class of all time–varying scalar quantizers (each one corresponding to a different sequence $s_1, \ldots, s_n$) subject to a code–length constraint $\sum_{i=1}^n R_{s_i} \leq nR$, or equivalently, $\sum_{s \in \mathcal{S}} n_s R_s \leq nR$, for a given pair (D, R).

In the notation of our generic model, here we have $U_i = X_i$, $v_i = s_i$, $i = 1, \ldots, n$, $q(u|v) = q(x|s) = q(x)$ independently of s, $f(u, v) = f(x, s) = -d(x, F_s(x))$, $E = -D$[6] and the excess distortion exponent is of the same form as before (see also [33]). Here, however, unlike the previous application examples, we have a degree of freedom to select the relative frequency of usage, $p(s)$, of each member of $\mathcal{F}$, i.e., the time–sharing protocol, but we also have the constraint $\sum_s p(s)R_s \leq R$.

From the statistical physics point of view, these additional ingredients mean that we have a freedom to select the number of particles in each subsystem (though the total number, n, is still fixed), and the additional constraint, $\sum_s p(s)R_s \leq R$, which is actually equivalent to the equality constraint $\sum_s p(s)R_s = R$ (in the interesting region of (R, D) pairs) can be viewed as an additional conservation law with respect to some other constant of motion, in addition to the energy (e.g., the

[6]One may prefer to redefine $f(x, s) = D_{\max} - d(x, F_s(x))$ and $E = D_{\max} - D$, where $D_{max} \stackrel{\Delta}{=} \max_{x,s} d(x, F_s(x))$, in order to work with non–negative quantities.

momentum), where in subsystem s, the (average) value of the corresponding physical quantity per particle is R_s.

While in [33], we have considered the problem of maximizing the rate function (the source coding exponent) of the excess distortion event $\sum_{i=1}^{n} d(X_i, F_{s_i}(X_i)) > nD$, a related objective (although somewhat less well motivated, but still interesting) is to minimize the rate function (or maximize the probability) of the small distortion event

$$\sum_{i=1}^{n} d(X_i, F_{s_i}(X_i)) < nD, \quad D < \sum_{(x,s)\in\mathcal{X}\times\mathcal{S}} q(x)p(s)d(x, F_s(x)).$$

In this case, the optimum performance is given by

$$F(R, D) = \max_{P\in\mathcal{P}(R)} \min_{\beta\geq 0} \left[\beta D + \sum_{s=1}^{S} p(s) \ln\left(\sum_{x\in\mathcal{X}} q(x)e^{-\beta d(x,F_s(x))}\right)\right],$$

where $\mathcal{P}(R)$ is the class of all probability distributions $P = \{p(s),\ s \in \mathcal{S}\}$ with $\sum_s p(s)R_s \leq R$. From the viewpoint of statistical physics, this corresponds to a situation where the various subsystems are allowed to interact, not only thermally, but also chemically, i.e., an exchange of particles is enabled in addition to the exchange of energy, and the maximization over $\mathcal{P}(R)$ (maximum entropy) is achieved when the chemical potentials of the various subsystems reach a balance. As the maximization over $P \in \mathcal{P}(R)$ subject to the constraint $\sum_s p(s)R_s \leq R$, for a given β, is a linear programming problem with one constraint (in addition to $\sum_s p(s) = 1$), then as was shown in [33], for each distortion level (or energy) D, the optimum $P \in \mathcal{P}(R)$ may be non–zero for at most two members of $\mathcal{S}$ only, which means that at most two subsystems are populated by particles in thermal and chemical equilibrium under the two conservation laws (of D and of R). However, the choice of these two members of $\mathcal{S}$ depends, in general, on D, which in turn depends on the temperature. Thus, when the system is heated gradually, certain *phase transitions* may occur, whenever there is a change in the choice of the two populated subsystems.

Finally, referring to comment no. 1 of Section 3, we should point out that here, in contrast to our discussion thus far, the difference between the ensemble of distinguishable particles and indistinguishable particles becomes critical since the factors $\{n_s!\}$ are no longer constant. Had we assumed indistinguishability, the normalized log–partition function would no longer be affine in P, thus the maximization over P would no longer be a linear programming problem, and the conclusion might have been different. In the source coding problem, the indistinguishable case

corresponds to a situation where the sequence of states s^n is chosen uniformly at random (with the decoder being informed of the result of the random selection, of course). In this case, the Chernoff bound corresponding to each composition $\{n_s,\ s \in \mathcal{S}\}$ of s^n should be weighed by the probability of this composition, which is $S^{-n}n!/\prod_s n_s!$. Now, each factor of $1/n_s!$ can be absorbed in the corresponding partition function $Z_s(\beta)$ of subsystem s, with the interpretation that in each subsystem the particles are now indistinguishable. The maximum over P would now correspond to the dominant contribution in this weighted average of Chernoff bounds. One can, of course, extend the discussion to any i.i.d. distribution on s^n, thus introducing additional bias and preferring some compositions over others.

Appendix

In this appendix, we outline another proof of Theorem 1 using a large deviations analysis approach. In particular, consider the large deviations event $\sum_{i=1}^n f(U_i, v_i) \leq nE$, as described in Section 2. Assuming that the relative frequencies $\{p(v)\}$ all stabilize as $n \to \infty$, let us compute the rate function $I(E)$ of the probability of this event in two different methods, where one would yield the left–hand side of (7) and the other would give the right–hand side of (7).

In the first method, we partition the sequence v^n according to its different letters. Specifically, let

$$E_v \stackrel{\Delta}{=} \frac{1}{n_v} \sum_{i:v_i=v} f(U_i, v),$$

where n_v is the number of occurrences of the symbol $v \in \mathcal{V}$ along v^n. Let $\mathcal{G}$ denote the set of all possible vector values that can be taken on by the vector $\bar{E} = \{E_v,\ v \in \mathcal{V}\}$. Now, obviously, $\sum_{i=1}^n f(U_i, v_i) \leq nE$ if and only if there exists a vector $\tilde{E} = \{\tilde{E}_v,\ v \in \mathcal{V}\} \in \mathcal{G}$ such that $E_v \leq \tilde{E}_v$ for all $v \in \mathcal{V}$ and $\sum_{v\in\mathcal{V}} p(v)\tilde{E}_v \leq E$. The "if" part follows from

$$\sum_{i=1}^n f(U_i, v_i) = n\sum_{v\in\mathcal{V}} p(v)E_v \leq n\sum_{v\in\mathcal{V}} p(v)\tilde{E}_v \leq nE.$$

The "only if" part follows by setting $\tilde{E}_v = E_v$ for all $v \in \mathcal{V}$. Therefore, denoting $\mathcal{H}_G(E) =$

$\mathcal{H}_0(E) \bigcap \mathcal{G}$ (where $\mathcal{H}_0(E)$ is defined as in Section 2), we have:

$$\begin{aligned}
\Pr\left\{\sum_{i=1}^{n} f(U_i, v_i) \le nE\right\} &= \Pr \bigcup_{\tilde{E}\in\mathcal{H}_G(E)} \left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v, \quad v \in \mathcal{V}\right\} \\
&\le \sum_{\tilde{E}\in\mathcal{H}_G(E)} \Pr\left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v, \quad v \in \mathcal{V}\right\} \\
&= \sum_{\tilde{E}\in\mathcal{H}_G(E)} \prod_{v\in\mathcal{V}} \Pr\left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v\right\} \\
&\le |\mathcal{H}_G(E)| \cdot \max_{\tilde{E}\in\mathcal{H}_G(E)} \prod_{v\in\mathcal{V}} \Pr\left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v\right\} \\
&\le |\mathcal{G}| \cdot \max_{\tilde{E}\in\mathcal{H}_G(E)} \prod_{v\in\mathcal{V}} \Pr\left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v\right\}, \qquad \text{(A.1)}
\end{aligned}$$

and on the other hand,

$$\begin{aligned}
\Pr\left\{\sum_{i=1}^{n} f(U_i, v_i) \le nE\right\} &= \Pr \bigcup_{\tilde{E}\in\mathcal{H}_G(E)} \left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v, \quad v \in \mathcal{V}\right\} \\
&\ge \max_{\tilde{E}\in\mathcal{H}_G(E)} \Pr\left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v, \quad v \in \mathcal{V}\right\} \\
&= \max_{\tilde{E}\in\mathcal{H}_G(E)} \prod_{v\in\mathcal{V}} \Pr\left\{\sum_{i:v_i=v} f(U_i, v) \le n_v \tilde{E}_v\right\}. \qquad \text{(A.2)}
\end{aligned}$$

At this point, the only gap between the upper bound (A.1) and the lower bound (A.2) is the factor $|\mathcal{G}|$. The number of different values that $\tilde{E}_v$ can take does not exceed the number of different type classes of sequences of length n_v over the alphabet $\mathcal{U}$, which is upper bounded by $(n_v + 1)^{|\mathcal{U}|-1}$.

Thus,

$$\begin{aligned}
|\mathcal{G}| &\leq \prod_{v\in\mathcal{V}}[n_v+1]^{|\mathcal{U}|-1} \\
&= \exp\left\{(|\mathcal{U}|-1)\sum_v \log(n_v+1)\right\} \\
&= \exp\left\{|\mathcal{V}|\cdot(|\mathcal{U}|-1)\sum_v \frac{1}{|\mathcal{V}|}\log(n_v+1)\right\} \\
&\leq \exp\left\{|\mathcal{V}|\cdot(|\mathcal{U}|-1)\log\left(\sum_v \frac{1}{|\mathcal{V}|}[n_v+1]\right)\right\} \\
&= \exp\left\{|\mathcal{V}|\cdot(|\mathcal{U}|-1)\log\left(\frac{n}{|\mathcal{V}|}+1\right)\right\} \\
&= \left(\frac{n}{|\mathcal{V}|}+1\right)^{|\mathcal{V}|\cdot(|\mathcal{U}|-1)}, \qquad \text{(A.3)}
\end{aligned}$$

and therefore $|\mathcal{G}|$ is only polynomial in n, and hence does not affect the exponential behavior. Now, each one of the terms $\Pr\{\sum_{i:v_i=v} f(U_i,v) \leq n_v\tilde{E}_v\}$ is bounded exponentially tightly by an individual Chernoff bound,

$$\exp\left\{n_v \min_{\beta\geq 0}\left[\beta\tilde{E}_v + \ln\left(\sum_u q(u|v)e^{-\beta f(u,v)}\right)\right]\right\},$$

and so, the dominant term of their product is of the exponential order of

$$\max_{\tilde{E}\in\mathcal{H}_G(E)}\sum_v p(v)\cdot\min_{\beta\geq 0}\left[\beta\tilde{E}_v+\ln\left(\sum_u q(u)e^{-\beta f(u,v)}\right)\right] = \max_{\tilde{E}\in\mathcal{H}_G(E)}\sum_v p(v)S_v(E_v).$$

Finally, as $n_v\to\infty$, the set $\mathcal{H}_G(E)$ becomes dense in the continuous set $\mathcal{H}_0(E)$, and by simple continuity arguments, the maximum over $\mathcal{H}_G(E)$ tends to the maximum over $\mathcal{H}_0(E)$.

The other method to evaluate the rate function $I(E)$ is as follows. Let ℓ be a fixed positive integer that divides n, and denote $\ell_v=\ell p(v)$, $v\in\mathcal{V}$ (assume that ℓ is chosen large enough that $\ell p(v)$ is well approximated by the closest integer with a very small relative error). Now, re–order the pairs $\{(U_i,v_i)\}$ (periodically), according to the following rule: Assuming, without loss of generality, that $\mathcal{V}=\{1,2,\ldots,|\mathcal{V}|\}$, the first $\ell_1=\ell p(1)$ symbol pairs of each ℓ–block of (u^n,v^n) are such that $v=1$, the next $\ell_2=\ell p(2)$ symbol pairs of each ℓ–block are such that $v=2$, and so on. In other words, each ℓ–block, $v_{(i-1)\ell+1}^{i\ell}=(v_{(i-1)\ell+1},v_{(i-1)\ell+2},\ldots,v_{i\ell})$, $i=1,2,\ldots,n/\ell$, consists of the same relative frequencies $\{p(v)\}$ as the entire sequence, v^n. Now, for the re–ordered sequence of pairs,

let us define $X_i = \sum_{t=(i-1)\ell+1}^{i\ell} f(U_t, v_t)$, $i = 1, 2, \ldots, n/\ell$. Obviously, $X_1, X_2, \ldots, X_{n/\ell}$ are i.i.d. and therefore the probability of the large deviations event $\{\sum_{i=1}^{n/\ell} X_i \leq \frac{n}{\ell} \cdot \ell E\}$ can be assessed exponentially tightly by the Chernoff bound as follows:

$$\begin{aligned}
& \exp\left\{\frac{n}{\ell} \cdot \min_{\beta \geq 0}\left[\beta \cdot \ell E + \ln\left(\sum_{u^\ell \in \mathcal{U}^\ell} q(u^\ell|v^\ell) \exp\left\{-\beta \sum_{i=1}^{\ell} f(u_i, v_i)\right\}\right)\right]\right\} \\
= & \exp\left\{\frac{n}{\ell} \cdot \min_{\beta \geq 0}\left[\beta \cdot \ell E + \ln\left(\prod_{v \in \mathcal{V}} \sum_{u^{\ell_v}} q(u^{\ell_v}|v^{\ell_v}) \exp\left\{-\beta \sum_{i=1}^{\ell_v} f(u_i, v)\right\}\right)\right]\right\} \\
= & \exp\left\{\frac{n}{\ell} \cdot \min_{\beta \geq 0}\left[\beta \cdot \ell E + \ln\left(\prod_{v \in \mathcal{V}} \left[\sum_{u \in \mathcal{U}} q(u|v) e^{-\beta f(u,v)}\right]^{\ell_v}\right)\right]\right\} \\
= & \exp\left\{\frac{n}{\ell} \cdot \min_{\beta \geq 0}\left[\beta \cdot \ell E + \ell \cdot \sum_{v \in \mathcal{V}} p(v) \ln\left(\sum_{u \in \mathcal{U}} q(u|v) e^{-\beta f(u,v)}\right)\right]\right\} \\
= & \exp\left\{n \cdot \min_{\beta \geq 0}\left[\beta E + \sum_{v \in \mathcal{V}} p(v) \ln\left(\sum_{u \in \mathcal{U}} q(u|v) e^{-\beta f(u,v)}\right)\right]\right\} \\
= & e^{n\bar{S}(E)}. \qquad \text{(A.4)}
\end{aligned}$$

Since both approaches yield exponentially tight evaluations of $I(E)$, they must be equal.

References

[1] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," arXiv:cond-mat/0106096, June 6, 2001.

[2] G. B. Bağci, "The physical meaning of Rényi relative entropies," arXiv:cond-mat/0703008v1, March 1, 2007.

[3] V. B. Balakirsky, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," *IEEE Trans. Inform. Theory*, vol. 41, no. 6, pp. 1889–1902, November 1995.

[4] T. Berger, *Rate distortion theory: a mathematical basis for data compression*, Prentice–Hall, Inc., Engelwood Cliffs, NJ, 1971.

[5] J. P. Burg, *Maximum entropy spectral analysis*, Ph.D. thesis, Department of Geophysics, Stanford University, Stanford, CA, 1975.

[6] M. Chiang and S. Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Trans. Inform. Theory*, vol. 50, no. 2, pp. 245–258, February 2004.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (second edition), John Wiley & Sons, Inc., New York, 2005.

[8] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*, New York: Academic, 1981.

[9] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 35–43, January 1995.

[10] R. S. Ellis, *Entropy, large deviations, and statistical mechanics*, Springer–Verlag, NY, 1985.

[11] R. S. Ellis, "The theory of large deviations and applications to statistical mechanics," lectures for international seminar on extreme events in complex dynamics, October 2006. Available on–line at: [http://www.math.umass.edu/∼rsellis/pdf-files/Dresden-lectures.pdf].

[12] R. M. Gray, *Source coding theory*, Kluwer Academic Publishers, 1990.

[13] R. M. Gray, A. H. Gray, G. Rebolledo, and J. E. Shore, "Rate distortionspeech coding with a minimum discrimination information distortion measure", *IEEE Trans. Inform. Theory*, vol. IT–27, no. 6, pp. 708–721, November 1981.

[14] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2325–2383, October 1998.

[15] D. Guo and S. Verdú, "Multiuser detection and statistical physics," in *Communications, Information and Network Security*, V. Bhargava, H. V. Poor, V. Tarokh, and S. Yoon, Eds., Chap. 13, pp. 229-277, Kluwer Academic Publishers, Norwell, Mass, USA, 2002.

[16] T. Hosaka and Y. Kabashima, "Statistical mechanical approach to error exponents of lossy data compression," *J. Physical Society of Japan*, vol. 74, no. 1, pp. 488–497, January 2005.

[17] Y. Iba, "The Nishimori line and Bayesian statistics," *J. Phys. A: Math. Gen.*, vol. 32, pp. 3875–3888, 1999.

[18] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev. A*, vol. 106, pp. 620–630, May 1957.

[19] E. T. Jaynes, "Information theory and statistical mechanics - II," *Phys. Rev. A*, vol. 108, pp. 171–190, October 1957.

[20] E. T. Jaynes, "On the rationale of maximum-entropy methods", *Proc. of the IEEE*, vol. 70, no. 9, pp. 939–952, September 1982.

[21] Y. Kabashima and T. Hosaka, "Statistical mechanics for source coding with a fidelity criterion," *Progress of Theoretical Physics*, Supplement no. 157, pp. 197–204, 2005.

[22] Y. Kabashima, K. Nakamura, and J. van Mourik, "Statistical mechanics of typical set decoding," *Physical Review E*, vol. 66, 2002.

[23] Y. Kabashima and D. Saad, "Statistical mechanics of error correcting codes," *Europhysics Letters*, vol. 45, no. 1, pp. 97–103, 1999.

[24] Y. Kabashima, N. Sazuka, K. Nakamura, and D. Saad, "Tighter decoding reliability bound for Gallager's error–correcting code," *Physical Review E*, vol. 64, pp. 046113-1–046113-4, 2001.

[25] I. Kanter and D. Saad, "Error–correcting codes that nearly saturate Shannon's bound," *Physical Review Letters*, vol. 83, no. 13, pp. 2660–2663, September 1999.

[26] C. Kittel, *Elementary statistical physics*, John Wiley & Sons, 1958.

[27] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Dev.*, vol. 5, pp. 183–191, 1961.

[28] A. Lapidoth, "Mismatched decoding and the multiple access channel,"Stanford Univ. Tech. Report, February 1994.

[29] A. Lapidoth and S. Shamai (Shitz), "A lower bound on the bit-error rate resulting from mismatched Viterbi decoding," Technical Report, CC Pub No. 163, Department of Electrical Engineering, Technion – I.I.T., August 1996.

[30] O. J. E. Maroney, "The (absence of a) relationship between thermodynamic and logical reversibility," arXiv:physics/0406137, June 27, 2004.

[31] D. McAllester, "A statistical mechanics approach to large deviations theorems," preprint, 2006. Available on-line at: [http://citeseer.ist.psu.edu/443261.html].

[32] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 6, pp. 1953–1967, November 1994.

[33] N. Merhav and I. Kontoyiannis, "Source coding exponents for zero–delay coding with finite memory," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 609–625, March 2003.

[34] M. Mézard and A. Montanari, *Constraint satisfaction networks in physics and computation*, draft, February 27, 2006. Available on–line at: [http://www.lptms.u-psud.fr/membres/mezard/].

[35] T. Mora and O. Rivoire, "Statistical mechanics of error exponents for error–correcting codes," arXiv:cond-mat/0606696, June 2006.

[36] T. Mutayama, "Statistical mechanics of the data compression theorem," *J. Phys. A: Math. Gen.*, vol. 35, pp. L95–L100, 2002.

[37] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 5, pp. 701–713, September 1982.

[38] H. Nishimori, "Spin glasses and information," *Physica A*, 384, pp. 94–99, 2007.

[39] Y. Oono, "Large deviation and statistical physics," *Progress of Theoretical Physics Supplement*, no. 99, pp. 165–205, 1989.

[40] S. Pereira and T. Weissman, "Denoising and filtering under the probability of excess loss criterion," *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1265–1281, April 2007.

[41] M. B. Plenio and V. Vitelli, "The physics of forgetting: Landauer's erasure principle and information theory," *Contemporary Physics*, vol. 42, no. 1, pp. 25–60, 2001.

[42] A. Procacci and B. Scoppola, "Statistical mechanics approach to coding theory," *J. of Statistical Physics*, vol. 96, nos. 3/4, pp. 907–912, 1999.

[43] I. Rojdestvenski and M. C. Cottman, "Mapping of statistical physics to information theory-with application to biological systems," *J. Theor. Biol.*, pp. 43–54, 2000.

[44] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.

[45] T. Shinzato, "Statistical physics and thermodynamics on large deviation," preprint. Available online at [http://www.sp.dis.titech.ac.jp/shinzato/LD.pdf].

[46] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT–26, no. 1, pp. 26–37, January 1980.

[47] N. Sourlas, "Spin–glass models as error–correcting codes," *Nature*, pp. 693–695, vol. 339, June 1989.

[48] N. Sourlas, "Spin glasses, error–correcting codes and finite–temperature decoding," *Europhysics Letters*, vol. 25, pp. 159–164, 1994.

[49] T. Tanaka, "Statistical mechanics of CDMA multiuser demodulation," *Europhysics Letters*, vol. 54, no. 4, pp. 540–546, 2001.

[50] T. Tanaka, "A statistical–mechanics approach to large–system analysis of CDMA multiuser detectors," *IEEE Trans. Inform. Theory*, vol. 48, no. 11, pp. 2888–2910, November 2002.

[51] M. J. Wainwright, T. S. Jaakkola, and S. S. Willsky, "A new class of upper bounds on the log partition function," *IEEE Trans. Inform. Theory*, vol. 51, no. 7, pp. 2313–2335, July 2005.

Neri Merhav - Biography

Neri Merhav (S'86–M'87–SM'93–F'99) was born in Haifa, Israel, on March 16, 1957. He received the B.Sc., M.Sc., and D.Sc. degrees from the Technion, Israel Institute of Technology, in 1982, 1985, and 1988, respectively, all in electrical engineering.

From 1988 to 1990 he was with AT&T Bell Laboratories, Murray Hill, NJ, USA. Since 1990 he has been with the Electrical Engineering Department of the Technion, where he is now the

Irving Shepard Professor. During 1994–2000 he was also serving as a consultant to the Hewlett–Packard Laboratories – Israel (HPL-I). His research interests include information theory, statistical communications, and statistical signal processing. He is especially interested in the areas of lossless/lossy source coding and prediction/filtering, relationships between information theory and statistics, detection, estimation, and Shannon Theory, including topics in joint source–channel coding, source/channel simulation, and coding with side information with applications to information hiding and watermarking systems.

Dr. Merhav was a co-recipient of the 1993 Paper Award of the IEEE Information Theory Society and he is a Fellow of the IEEE since 1999. He also received the 1994 American Technion Society Award for Academic Excellence and the 2002 Technion Henry Taub Prize for Excellence in Research. From 1996 until 1999 he served as an Associate Editor for Source Coding to the IEEE Transactions on Information Theory. He also served as a co–chairman of the Program Committee of the 2001 IEEE International Symposium on Information Theory. He is currently on the Editorial Board of Foundations and Trends in Communications and Information Theory.